Available online at www.sciencedirect.com
SCIENCE DIRECT®
Mutation Research 549 (2002) 241–253
Fundamental and Molecular Mechanisms of Mutagenesis
www.elsevier.com/locate/molmut
Community address: www.elsevier.com/locate/mutres

# Development of public toxicogenomics software for microarray data management and analysis

Weida Tong [a,*], Stephen Harris [b], Xiaoxi Cao [b], Hong Fang [b], Leming Shi [a], Hongmei Sun [b], James Fuscoe [c], Angela Harris [d], Huixiao Hong [b], Qian Xie [b], Roger Perkins [b], Dan Casciano [e]

[a] *Center for Toxicoinformatics, Division of Biometry and Risk Assessment, NCTR, 3900 NCTR Road, HFT-020, Jefferson, AR 72079, USA*
[b] *Northrop Grumman Information Technology, Jefferson, AR 72079, USA*
[c] *Center for Functional Genomics, Division of Reproductive and Genetic Toxicology, NCTR, FDA, Jefferson, AR 72079, USA*
[d] *Center for Hepatotoxicity, NCTR, FDA, Jefferson, AR 72079, USA*
[e] *Office of Director, NCTR, FDA, Jefferson, AR 72079, USA*

Received 9 October 2003; received in revised form 19 December 2003; accepted 22 December 2003

**Abstract**

A robust bioinformatics capability is widely acknowledged as central to realizing the promises of toxicogenomics. Successful application of toxicogenomic approaches, such as DNA microarray, inextricably relies on appropriate data management, the ability to extract knowledge from massive amounts of data and the availability of functional information for data interpretation. At the FDA's National Center for Toxicological Research (NCTR), we are developing a public microarray data management and analysis software, called ArrayTrack. ArrayTrack is Minimum Information About a Microarray Experiment (MIAME) supportive for storing both microarray data and experiment parameters associated with a toxicogenomics study. A quality control mechanism is implemented to assure the fidelity of entered expression data. ArrayTrack also provides a rich collection of functional information about genes, proteins and pathways drawn from various public biological databases for facilitating data interpretation. In addition, several data analysis and visualization tools are available with ArrayTrack, and more tools will be available in the next released version. Importantly, gene expression data, functional information and analysis methods are fully integrated so that the data analysis and interpretation process is simplified and enhanced. ArrayTrack is publicly available online and the prospective user can also request a local installation version by contacting the authors.


*Keywords:* Toxicogenomics software; DNA microarray; Database

## 1. Introduction

The use of "omics" technologies to assess the gene/protein expression changes in chemical- and/or environment-induced toxicity, with emphasis on determination of corresponding gene/protein functions, pathways, and regulatory networks, are driving the emergence of the new research field of toxicogenomics [1]. DNA microarray is one of the main technological advances that has revolutionized both the theory and practice of addressing toxicological questions at the molecular level [2–4].

* Corresponding author. Tel.: +1-870-543-7142; fax: +1-870-543-7662.
*E-mail address:* wtong@nctr.fda.gov (W. Tong).

0027-5107/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.mrfmmm.2003.12.024

A DNA microarray experiment proceeds through hypothesis, experimental design and gene expression measurement in a manner similar to a conventional toxicology study. The amount and nature of the data associated with a microarray experiment, however, impose unique challenges requiring bioinformatics support. There are three major bioinformatics issues important for the success of the experiment:

- *Data management*. This step acquires essential information associated with a microarray experiment. A microarray experiment involves multiple steps and the data in each step needs to be appropriately managed, annotated, and most importantly centralized. First, this is convenient for the subsequent data analysis that normally requires a multidisciplinary group of scientists to access the same dataset. Second, because gene annotation is continuously updated in the public domain, the analyzed data need to be re-examined periodically. Lastly, given that the analysis methods are rapidly evolving, a well-managed and annotated dataset can be easily reanalyzed.
- *Data analysis*. A single experiment can produce a large amount of data and a formidable analysis undertaking. Normally, the immensity of data analysis scales directly with the complexity of the experiment, such as the number of technical and biological replicates, and temporal and dose response parameters. The ability to search, filter, and apply mathematical and statistical operations and graphically visualize data quickly with an intuitive user interface is crucial to the laborious process.
- *Data interpretation*. Experiment interpretation is a highly contextual process in light of known and unknown functions of genes, proteins and pathways. The inherent noise in microarray data and a plethora of potential sources of variability inevitably complicate and possibly confound interpretation. Efficient and effective interpretation demands that relevant knowledge residing in public sources for gene annotation, protein function and pathways are readily available and integrated with the data analysis process.

At the FDA's National Center for Toxicological Research (NCTR/FDA), the microarray core facility using a two-channel microarray platform has been established that utilizes standardized experimental procedures to conduct toxicogenomic research in a collaborative environment. Correspondingly, a microarray data management, analysis and interpretation software system, ArrayTrack, has been developed to address the bioinformatics challenges associated with microarray experiment [5].

The distinct features of ArrayTrack compared to other database software from public and commercial vendors are: (1) a quality control mechanism is implemented to assure the fidelity of entered expression data; (2) a rich collection of functional information about genes, proteins and pathways is available for facilitating data interpretation; (3) gene expression data, functional information and analysis methods are integrated so that the data analysis and interpretation process is simplified and enhanced; (4) conventional toxicological data and gene expression data are cross-linked to facilitate investigation of toxicity at the molecular level; and (5) the system is easy to be extended to accommodate other types of "omics" data (e.g., proteomic and metabonomic data) for the "systeomic" research.

At the time of this writing, the ArrayTrack version 2.01 can be accessed through http://www.edkb.fda.gov/webstart/arraytrack (http://www.weblaunch.nctr.fda.gov/jnlp/arraytrack for FDA users). Prospective users also can acquire a free distribution of the software by contacting the authors. In this paper, the main features of ArrayTrack are described with emphasis on the practical issues and rationale behind the software development.

## 2. Methods

ArrayTrack is a client–server system. The ORACLE server stores and integrates in-house omics data and data from public resources about genes, proteins and pathways. The JAVA language was used to construct the entire user interface, query mechanism, and data visualization and analysis tools. The use of JAVA ensures portability of ArrayTrack to all major computer operating systems, as well as enabling easy web-deployment. The client–server connection is realized through JDBC (JAVA Database Connectivity). The use of JDBC makes it easy for ArrayTrack to use other relational databases for backend storage, since dependency on ORACLE is minimal.

ArrayTrack has a modular architecture. Each module for each application is constructed independently, such that existing or new capabilities can be enhanced, changed or added in accordance with priorities and evolving experimental progress. Thus, ArrayTrack is under continuous development and updating. Although ArrayTrack is 100% Java, integration with non-Java applications can readily be made through socket-based communication on a local machine, provided the other application can be scripted or if small programming changes in the other application can be made.

ArrayTrack is implemented using Java Webstart technology, which allows installation through a single web link with updates of the software performed automatically whenever the application is run. The software has been fully tested on Microsoft Windows (98/NT/2000/XP) and Unix platforms (including Linux).

## 3. Results and discussion

### 3.1. Managing toxicogenomics data

ArrayTrack supports the Minimum Information About a Microarray Experiment (MIAME) guideline. MIAME defines essential information for a microarray experiment that enables the results to be interpretable and the experiment to be reproducible [6]. Currently, a number of journals, including Nature, the Nature group of journals, Cell, The Lancet, EMBO and Toxicology Pathology, requires an accession number from the public microarray databases developed based on the MIAME guidelines to be supplied at or before acceptance of publication [7].

Microarray information for a toxicogenomic study can be input and viewed/edited through a comprehensive data submission form in ArrayTrack (Fig. 1A). The form contains three sections:

- *Experiment design*. An experiment's hypothesis and the associated experimental protocols are input in this section. The owner of the experiment can assign "read and/or write" privilege for experiment information and results with collaborators and others. A list of genes anticipated to be significant as a consequence of the experiment hypothesis and design is also input in this section, which may serve as a toxicity-specific expression signature and be used for cross-experiment comparisons.
- *Hybridization and data*. The description of the hybridization process and the raw data are input in this section. Both the raw images and the associated numerical intensities are stored. ArrayTrack supports both one- and two-channel microarray experiments including Affymetrix data.
- *Sample*. An accurate description of animals and treatments is an essential task of toxicogenomics research for establishing association of genomic data with phenotype. The critical information associated with the samples (normally associated with animal tissue in toxicogenomic research) are input in this section. In the future, a version storing more extensive information about samples will be implemented in accordance with the MIAME/Tox guidance (http://www.mged.org) that will expand the original MIAME proposal to encompass additionally required information for toxicogenomic experiments.

It is common that hypothesis generation, hybridization experiment and sample preparation might be conducted by different groups of people in an organization that, specifically, has a microarray core facility. This specific design of the form is advantageous in such a collaborative environment, where information can be separately entered into each section by different scientists.

### 3.2. Assuring quality of expression data

A database is only useful when the quality of entered data is indexed. Only a validated database can be a rich resource for cross-experiment and cross-platform comparisons to derive toxicity-specific signatures. Microarray experimentation has become one of the fastest-growing methods, and has led to a broad diversity of microarray databases in both the public and commercial domains [8–10]. Although the importance of quality control (QC) is generally understood, there is little QC practice in the existing microarray databases.

We implemented an approach for the QC of two-channel microarray data (Fig. 1B). The QC page summarizes the most relevant information about a slide into one interface for a Pass/Fail/Review call.

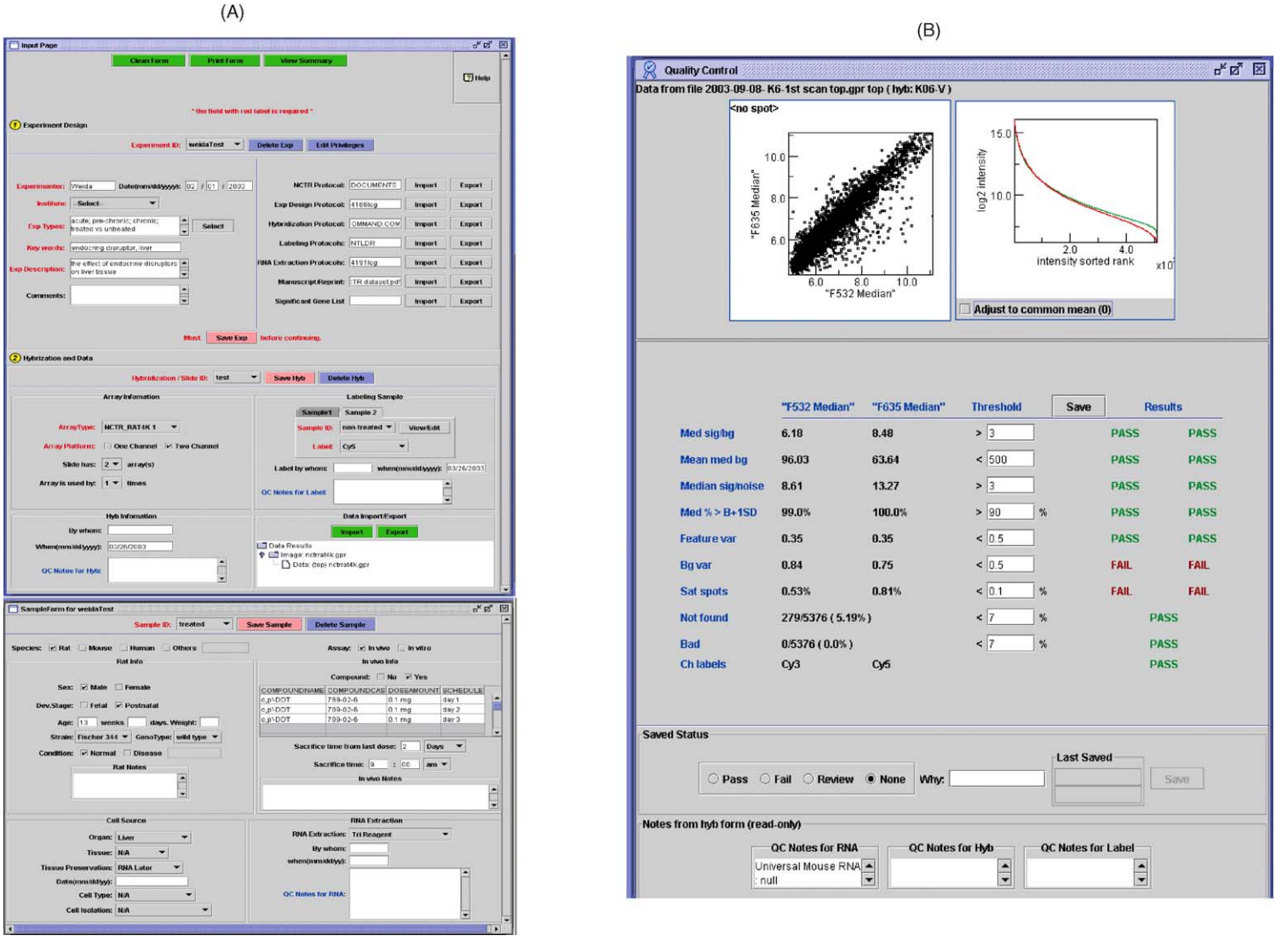



Fig. 1. ArrayTrack data submission form (A) and quality control panel (B).

The user can determine the quality of individual microarray results through visualizing data, applying statistical measures and viewing experimental annotation. Statistical measures are provided to assess the quality of a hybridization result based on the raw expression data, including signal-to-noise and signal-to-background ratio, the percentage of non-hybridized and saturated spots, etc. The experimental annotations associated with the processes of hybridization, RNA extraction and labeling are also available to the end-user. Additionally, a scatter plot of Cy3 versus Cy5 (or an M–A plot) together with the rank intensity plot (RIP) of both channels is available for visual inspection. (The functions of the scatter plot, M–A plot and RIP are discussed in the section "Analyzing and visualizing expression data.") The plots and statistics are dynamically linked. Users are able to examine the quality of a slide based on a specific set of genes or the entire list of genes. Being able to examine a subset of genes is useful in the QC process because the user can determine quickly the quality of a slide based on selected genes, such as housekeeping genes, spike or positive/negative control genes. Importantly, each QC decision is recorded in the database, permitting later development of a supervised learning model that relates calculated statistics with QC decisions; such a model, when automated, could eliminate tedious human efforts and provide standardized and unbiased QC decisions for large numbers of experiments.

### 3.3. Aggregating functional information about genes, proteins and pathways

The public domain has a rich and diverse collection of biological databases that provide functional information useful for microarray experiment interpretation and associated knowledge discovery [11]. Some public databases are undergoing rapid update and expansion. For example, the gene ontology (GO) consortium [12] maintains a controlled vocabulary database of functional descriptions for genes in terms of three functional categories, biological process, molecular function and cellular component. As shown in Table 1, the total number of functional descriptions (GO terms) as well as the numbers in the individual categories have almost doubled within a period of 6 months (January–July, 2003).

Table 1
Comparison of the number of terms in gene ontology between 24 January 2003 and 24 July 2003

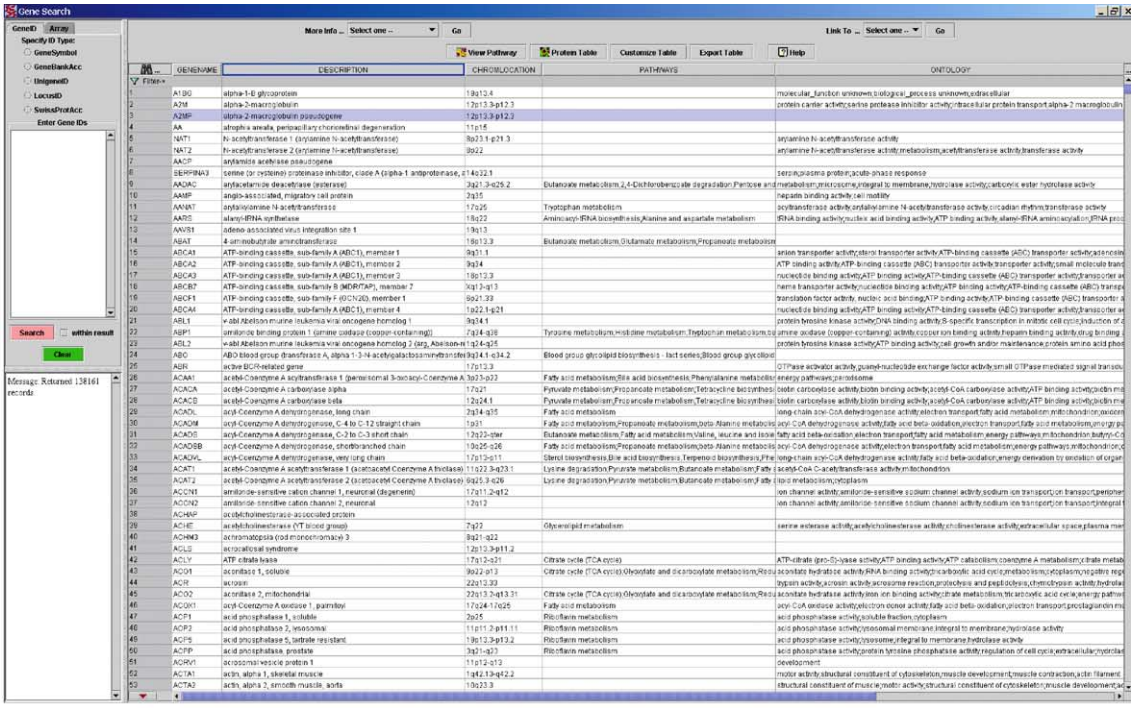
| | 24 January 2003 | 24 July 2003 |
|---|---|---|
| Total number of GO terms | 46199 | 80972 |
| The number of terms for the category of | | |
| Biological process | 30188 | 56741 |
| Molecular function | 37018 | 66225 |
| Cellular component | 22371 | 38547 |

We developed several ORACLE databases to mirror the contents of GenBank, SWISS-PROT, LocusLink, Kyoto Encyclopedia of Genes and Genomes (KEGG), GO and others. GenBank [13] contains sequence data (coding, genomic, EST, and synthetic) with basic annotation, while SWISS-PROT [14] is a protein sequence database of low redundancy with high levels of annotation. LocusLink [15] offers a simple query interface to retrieve information about human genes and some non-gene loci, and also provides direct connections to related information available from other resources. One of the major components of KEGG [15] provides information on metabolic and regulatory pathways. In order to keep current, we update the content of our mirrored databases using scripts every 2 weeks. Importantly, we extract the functional information from these databases to construct three enriched libraries, GeneLib (Fig. 2A), ProteinLib and PathwayLib. As the names suggest, these three libraries concentrate functional information on genes, proteins and pathways, respectively [5]. The user can quickly identify the functional information for a set of significant genes derived from analysis by searching these libraries.

### 3.4. Concerning the representation of genes on a chip

The difficulties associated with producing cDNA microarray in terms of purifying PCR products and managing the cDNA banks have led to wide use of short oligonucleotides to represent the desirable genes on a chip. For example, Affymetrix (Affymetrix, Santa Clara, CA) uses 25-mer, MWG (MWG Biotech, High Point, NC) uses 50-mer, Agilent (Agilent Technologies, Palo Alto, CA) uses 60-mer, Operon (QIAGEN Operon, Alemeda, CA) uses 70-mer, and Clontech (BD Biosciences, Palo Alto, CA) uses 80-mer for their oligonucleotide microarray fabrication. The potential

(A)




(B)

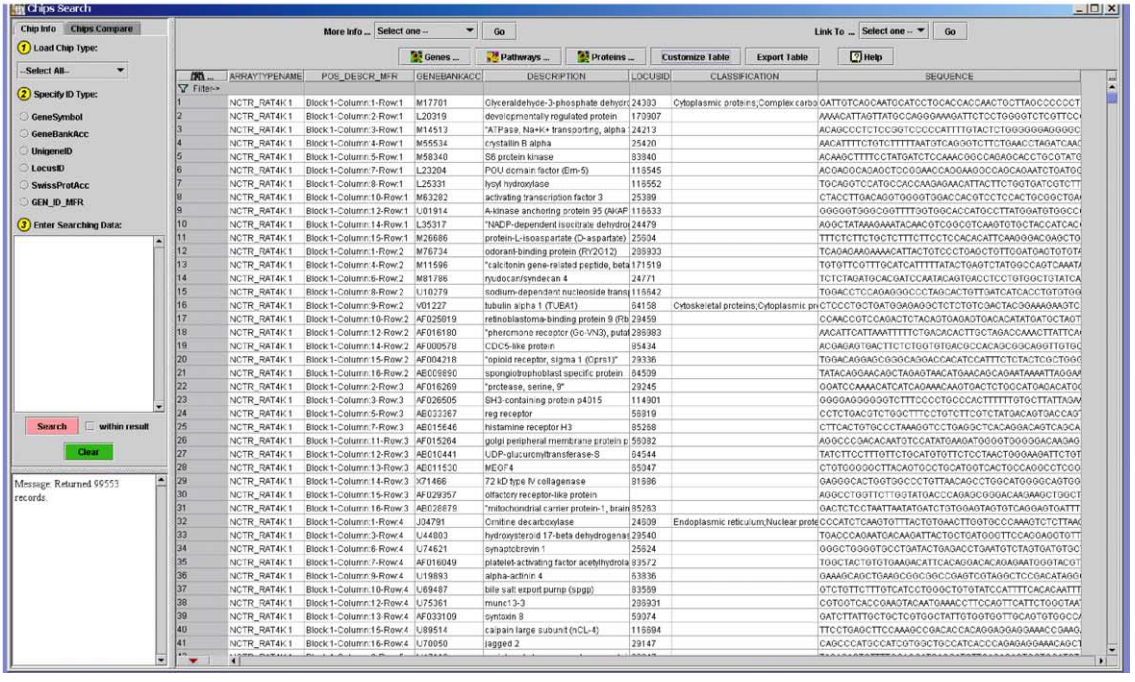



Fig. 2. ArrayTrack GeneLib (A) and ChipLib (B).

cross-hybridization of oligonucleotide arrays has been a general concern and there are still different opinions on the best length of oligonucleotide probes. More data are needed to find a probe length that best balances specificity and sensitivity.

Given the debate on which length of probes provides better indication (representation) of a gene and has less cross-hybridization, storing sequence of probes is useful for cross-platform comparison. We have developed ChipLib (Fig. 2B) that contains all functional information provided by the manufactures for the probes on a chip, including the sequence. Moreover, since understanding the function and biological characteristics of the probes (genes) presented on a microarray could be essential for interpretation of microarray results, genes represented on the array are also directly linked with the GeneLib, ProteinLib and PathwayLib for facilitating biological interpretation of experiment results.

### 3.5. Linking expression data with data from conventional toxicology study

The combination of expression data with more traditional toxicology data and chemical structure information to determine phenotypic responses to toxicants at the mechanistic level is one of the important research goals of toxicoinformatics. Thus, an additional library, ToxicantLib, is being developed to provide linkage between toxicological data and the expression data. The ToxicantLib explicitly contains chemical structure together with toxicological endpoints. Since chemicals with similar structures are likely to exhibit similar biological (or toxicological) activities [16], we are also implementing an algorithm for assessing structure similarity of chemicals and exploring structure–toxicity relationship based on the substructure features and physicochemical properties derived from the structure. ToxicantLib has been initially populated with data from our Endocrine Disruptor Knowledge Base (EDKB) [17] and the Carcinogenicity Potency Database (CPDB) [18].

### 3.6. Analyzing and visualizing expression data

Several tools for data normalization, analysis and visualization are implemented in ArrayTrack. The raw expression data in ArrayTrack can be manually or automatically processed using two global normalization approaches, total intensity normalization [19] and log ratio mean scale normalization [20].

RIP sorts intensities of genes in a descending order along the *y*-axis, and each gene is given an ordinal number along the *x*-axis to reflect its relative position on a chip [21]. The green curve represents the cy3-labeled samples and the red curve represents the cy5-labeled sample. The shape of the curves characterizes the general properties of the expression data. Well-balanced two-channel microarray data should show a superimposed or parallel distribution of the green and red lines (Fig. 3A). The crossover of the green and red lines shown in Fig. 3B indicates the unbalanced bias between the two channels. Thus, RIP can give a general impression about the quality of data.

The ScatterPlot Viewer provides the pair-wise plotting of Cy3 versus Cy5 for two-channel microarray

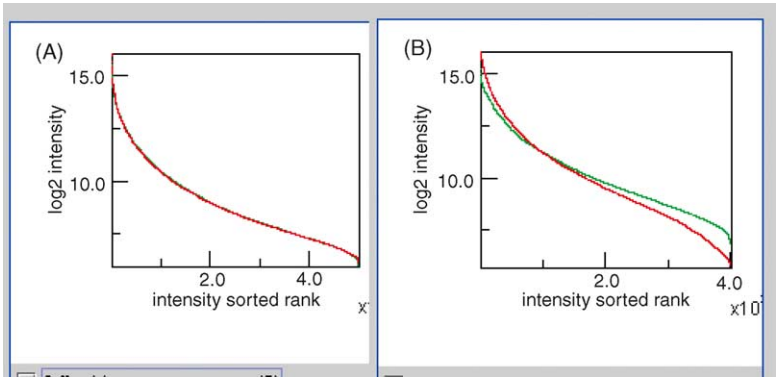



Fig. 3. Rank intensity plot for a balanced (A) and an unbalanced (B) two-channel array.

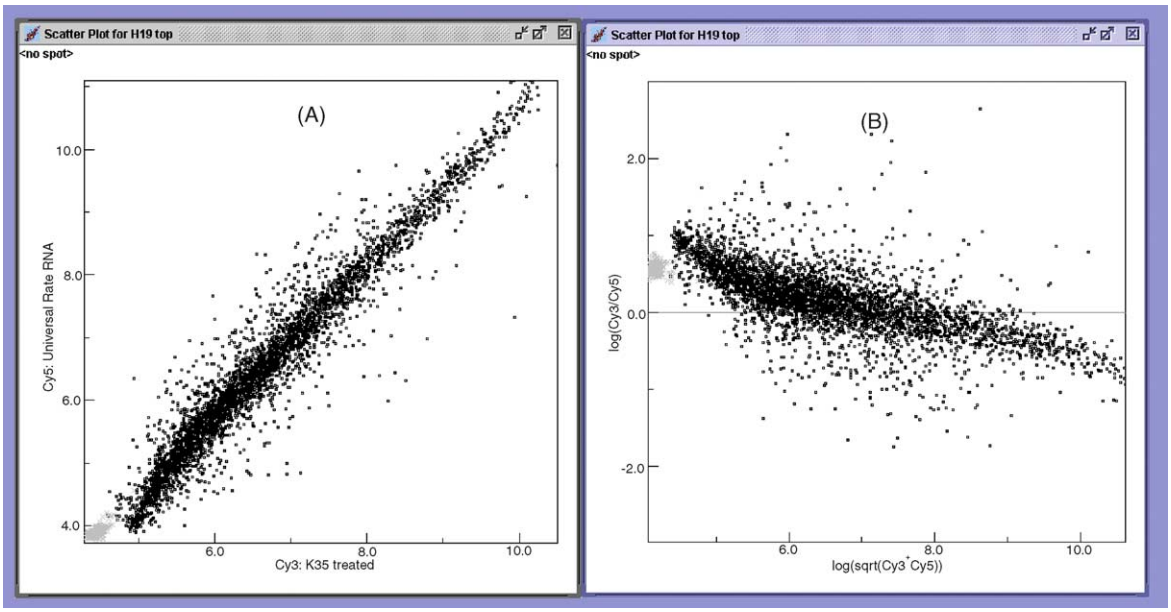



Fig. 4. Scatter plot (A) and M–A plot (B) for the data shown in Fig. 2B.

data, and also plots expression data of one sample versus another for one-channel microarray data. Since the display of a single-slide expression data using the scatter plot does not fully reflect the concordance and quality of the data, we also provide an M–A plot [22], where the log intensity ratio $M = \log_2(\text{Cy5}/\text{Cy3})$ is plotted against the mean log intensity $\text{A} = 0.5\log_2(\text{Cy3} \times \text{Cy5})$. A comparison of the scatter plot with the M–A plot is given for the same single-slide expression data used in Fig. 3B. Although the scatter plot (Fig. 4A) still shows a good concordance of two channels, the unbalanced nature of the data is revealed in the M–A plot (Fig. 4B), where the data are not parallel along the *x*-axis.

The VirtualArray Viewer displays expression data in the format of the original array image (Fig. 5). This function reconstructs the original array image based on either the raw or normalized expression data and provides a visual representation of data for further exploration, analysis and interpretation. For example, there are two sliding controls on the top of the image for filtering out unwanted spots. The upper sliding control is used to eliminate spots whose expression fold change is less than a predefined criterion (e.g., two-fold). The other sliding control is used to eliminate spots for which the intensity of both Cy3 and Cy5 channels falls below the selected threshold. The user can also search the image to identify the position of a selected list of genes, which could be useful to examine the reliability of the differentially expressed gene list. For example, care must be taken if most significant genes are located in a specific block, which usually indicates a flaw of that block.

The BarChart Viewer compares the expression level of a gene across the array data within a single experiment or across multiple experiments and/or platforms (Fig. 6). Each bar is associated with a particular array and the height of a bar indicates the expression level (fold-change for the two-channel array and intensity for one-channel array) of a gene that can be represented by the data of either before or after normalization. This function can be useful for examining the dose–response relationship and time-dependent pattern of a specific gene.

Given the broad availability and selection of microarray data analysis tools in both commercial and public domains, we are focusing on developing interfaces to provide interoperability between ArrayTrack and other analysis software. In the next release

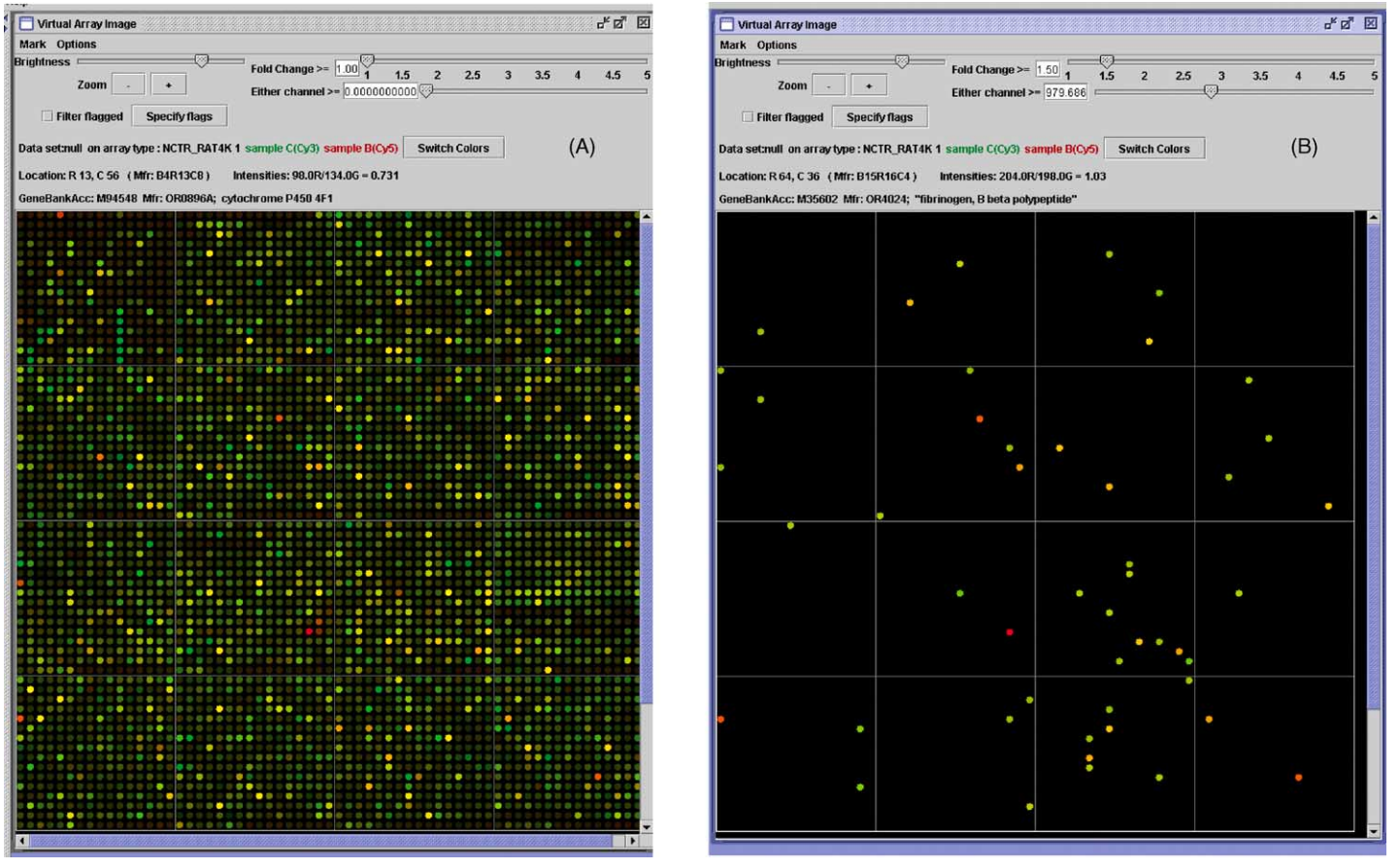



Fig. 5. VirtulArray Viewer. It shows the images of before (A) and after (B) filtering using two sliding controls.

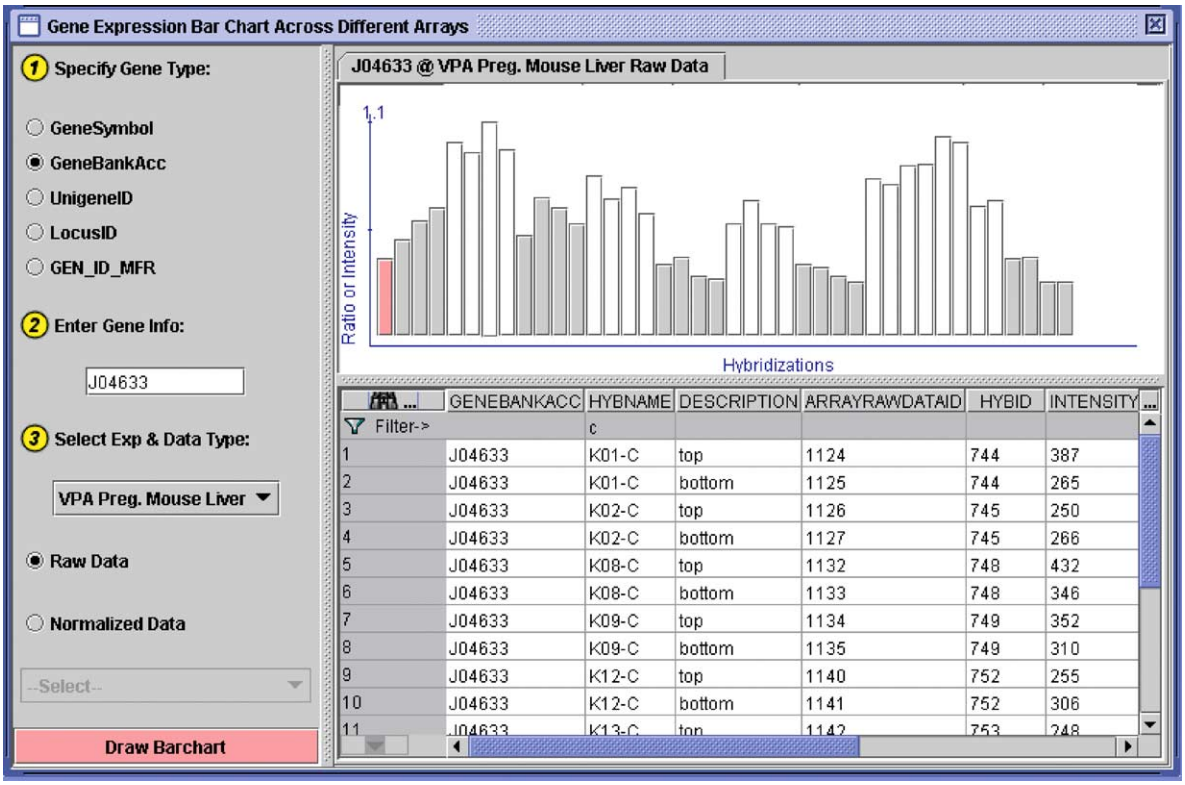



Fig. 6. BarChart Viewer. It shows that the gene (heat shock protein, Hspca, GenBank Acc# J04633) is over-expressed in the drug-treated mice compared to the control mice for total of 40 array data.

of the software, the full integration of ArrayTrack with TASS from Chipscreen (http://www.chipscreen.com/chinese_gb/chipservice/tass.htm) will allow users to access a number of data mining and data analysis functions, including hierarchical cluster analysis, principal component analysis, self-organizing maps and support vector machines (Fig. 7).

### *3.7. Integrating analysis with functional information for biological interpretation*

The primary emphasis of ArrayTrack is the direct linking of analysis results with functional information for facilitating the interaction between the choice of analysis methods and the biological relevance of analysis results. Using ArrayTrack, the user can select an analysis method and apply it to the stored microarray data, and the analysis results can be directly linked to gene, protein and pathway information in the libraries. Additionally, ArrayTrack also allows analysis results to be directly linked with other public databases.

One major benefit derived from the integration of analysis methods with the functional information is the immediate feedback that can be given to the biologists so that the biological interpretation can be rapidly investigated. This, in turn, will lead to the selection of the optimal analysis method. The integrating process is necessary given that there are many choices of methods available for analyzing microarray data and, unfortunately, it is often difficulty to determine the best choice. For example, even for the well-defined hierarchical clustering analysis, many different options are available and they may produce different results for the same data sets. In such a situation, the choice of the analysis method is dependent on the biological relevance of the results derived from the method. The integration of the analysis method with functional information will improve the ability and reduce time for data interpretation.

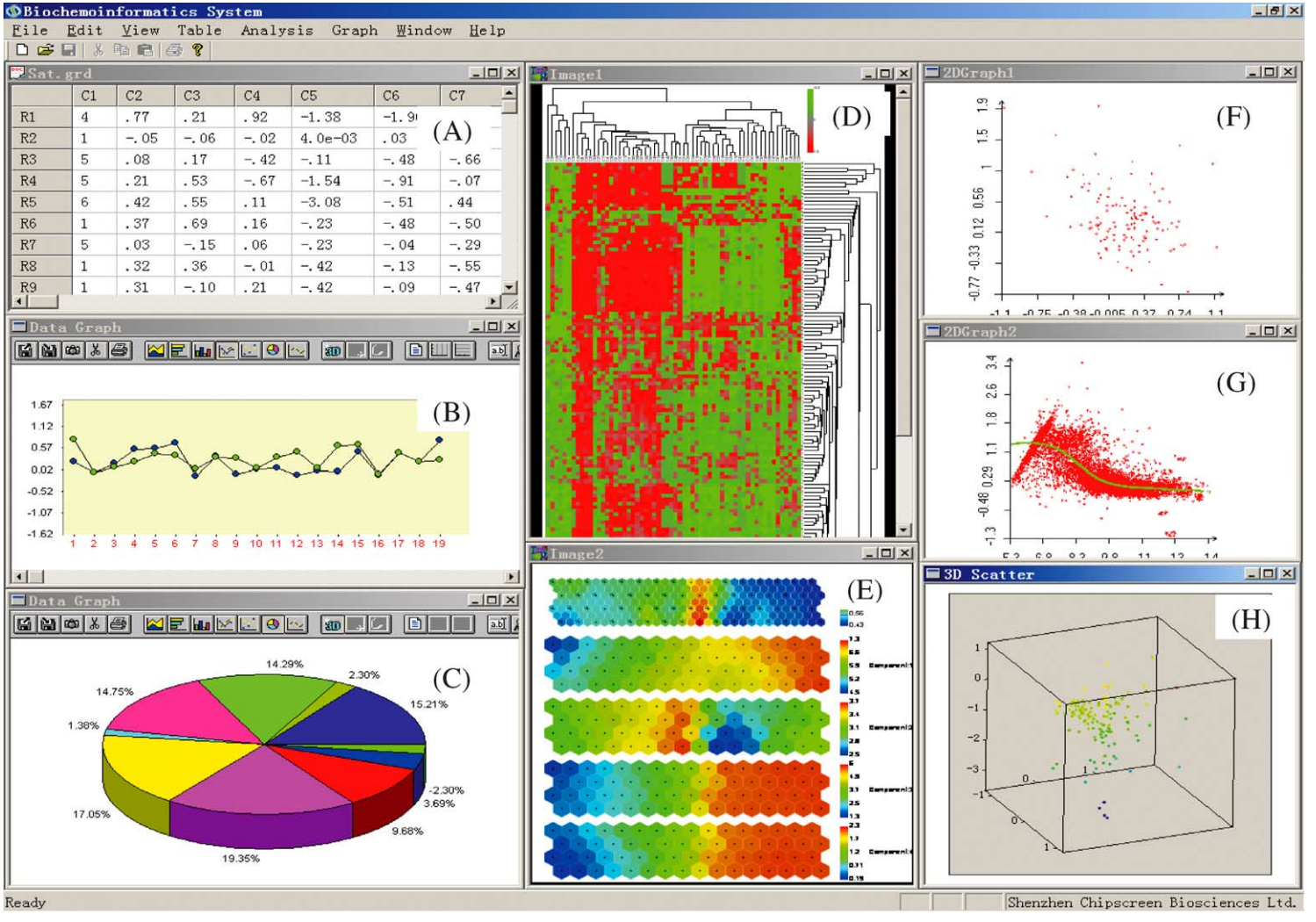



Fig. 7. Preview of analysis functions in the next release of ArrayTrack: (A) a spreadsheet of gene expression data; (B) time-course view of gene expression data for two genes; (C) Pie chart; (D) two-way hierarchical clustering and color image display of expression data; (E) self-organizing map; (F) 2D scatter plot; (G) LOWESS fitting; and (H) 3D scatter plot view of PCA results.

## 4. Future directions and challenges

DNA microarray has been enjoying steady growth of impact on toxicology and the trend is certain to continue. This paradigm shift in toxicology is largely facilitated by unprecedented advancement in bioinformatics and other informatics-related fields. However, DNA microarray technology is still rapidly evolving, in part, owing to the fact that large variability is still observed for most microarray platforms and it is difficult to generate comparable results from different platforms. The current technology often generates more questions than answers for application in toxicogenomics and specifically in regulation. It is proposed that more reliable conclusions may be reached by integrating gene expression data with other omics data, such as data from proteomics and metabolomics research, as well as data from traditional toxicological studies and chemical structure information. The linkage of these types of information presents challenges and opportunities to develop a comprehensive and robust toxicogenomics software system to accommodate diverse data from various sources. This "systeomic" approach will facilitate scientific discovery and productivity via effective management of diverse toxicological data and knowledge at different levels of biological complexity, which will lead to more fully understanding toxicity at the mechanistic level.

To effectively meet the challenges of future toxicogenomic research, we have been extending the scope of ArrayTrack for the purpose of fully integrating genomic, proteomic, and metabonomic data with data from the public repositories, as well as conventional in vitro and in vivo toxicology data at NCTR/FDA. This extensive version of ArrayTrack, called Toxicoinformatics Integrated System (TIS) will serve for toxicogenomics as a general repository for diverse data sources, supporting broad data mining and meta-analysis activities, as well as the development of robust and validated predictive toxicology systems.

## Acknowledgements

The authors gratefully acknowledge Drs. Robert Delongchamp and Varsha Desai at NCTR, who provided encouragement and insightful suggestions for the development of ArrayTrack.

## References

[1] C.W. Schmidt, Toxicogenomics: an emerging discipline, Environ. Health Perspect. 110 (2002) A750–A755.
[2] C.A. Afshari, E.F. Nuwaysir, J.C. Barrett, Application of complementary DNA microarray technology to carcinogen identification, toxicology, and drug safety evaluation, Cancer Res. 59 (1999) 4759–4760.
[3] E.F. Nuwaysir, M. Bittner, J. Trent, J.C. Barrett, C.A. Afshari, Microarrays and toxicology: the advent of toxicogenomics, Mol. Carcinog. 24 (1999) 153–159.
[4] H.K. Hamadeh, R.P. Amin, R.S. Paules, C.A. Afshari, An overview of toxicogenomics, Curr. Issues Mol. Biol. 4 (2002) 45–56.
[5] W. Tong, X. Cao, S. Harris, H. Sun, H. Fang, C.J. Fuscoe, H. Hong, Q. Xie, R. Perkins, L. Shi, D. Casciano, ArrayTrack-supporting toxicoinformatic research at the FDA's National Center for Toxicological research (NCTR), EHP Toxicogenomics 111 (2003) 1819–1826.
[6] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, F.C. Holstege, I.F. Kim, V. Markowitz, J.C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, M. Vingron, Minimum information about a microarray experiment (MIAME) toward standards for microarray data, Nat. Genet. 29 (2001) 365–371.
[7] Microarray standards at last, Nature 419 (2002) 323.
[8] M. Gardiner-Garden, T.G. Littlejohn, A comparison of microarray databases, Brief Bioinform. 2 (2001) 143–158.
[9] P. Anderle, M. Duval, S. Draghici, A. Kuklin, T.G. Littlejohn, J.F. Medrano, D. Vilanova, M.A. Roberts, Gene expression databases and data mining, Biotechniques Suppl. (2003) 36–44.
[10] S. Dudoit, R.C. Gentleman, J. Quackenbush, Open source software for the analysis of microarray data, Biotechniques Suppl. (2003) 45–51.
[11] A.D. Baxevanis, The Molecular Biology Database collection: 2003 update, Nucl. Acids Res. 31 (2003) 1–12.
[12] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The gene ontology consortium, Nat. Genet. 25 (2000) 25–29.
[13] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, D.L. Wheeler, GenBank, Nucl. Acids Res. 31 (2003) 23–27.
[14] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, M. Schneider, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, Nucl. Acids Res. 31 (2003) 365–370.

[15] M. Kanehisa, The KEGG database, Novartis Found. Symp. 247 (2002) 91–101, discussion 101–103, 119–128, 244–252.
[16] M. Johnson, G.M. Maggiora, Concepts and Applications of Molecular Similarity, Wiley, New York, 1990.
[17] W. Tong, R. Perkins, H. Fang, H. Hong, Q. Xie, S.W. Branham, D.M. Sheehan, J.F. Anson, Development of quantitative structure-activity relationships (QSARs) and their use for priority setting in the testing strategy of endocrine disruptors, Regul. Res. Perspect. 1 (2002) 1–16.
[18] L.S. Gold, E. Zeiger, Handbook of Carcinogenic Potency and Genotoxicity Databases, CRC Press, 1997.
[19] J. Quackenbush, Microarray data normalization and transformation, Nat. Genet. 32 (Suppl) (2002) 496–501.
[20] M.R. Fielden, R.G. Halgren, E. Dere, T.R. Zacharewski, GP3: GenePix post-processing program for automated analysis of raw microarray data, Bioinformatics 18 (2002) 771–773.
[21] T.C. Kroll, S. Wolfl, Ranking: a closer look on globalisation methods for normalisation of gene expression arrays, Nucl. Acids Res. 30 (2002) e50.
[22] Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, T.P. Speed, Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, Nucl. Acids Res. 30 (2002) 15.